Data Analysis Projects
MoBi 4. FS - SS2020

Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

Concept

Medizinische Fakultät Heidelberg

- **Project-oriented teaching**: provide hands-on experience with data analysis and programming

- **Goals**
  - experience **data analysis challenges** on real datasets related to research question
  - experience team work, also outside of your team!
  - learn to use modern data analysis tools: R / Python / markdown / notebooks / github

Learning by doing!

Research topics / projects

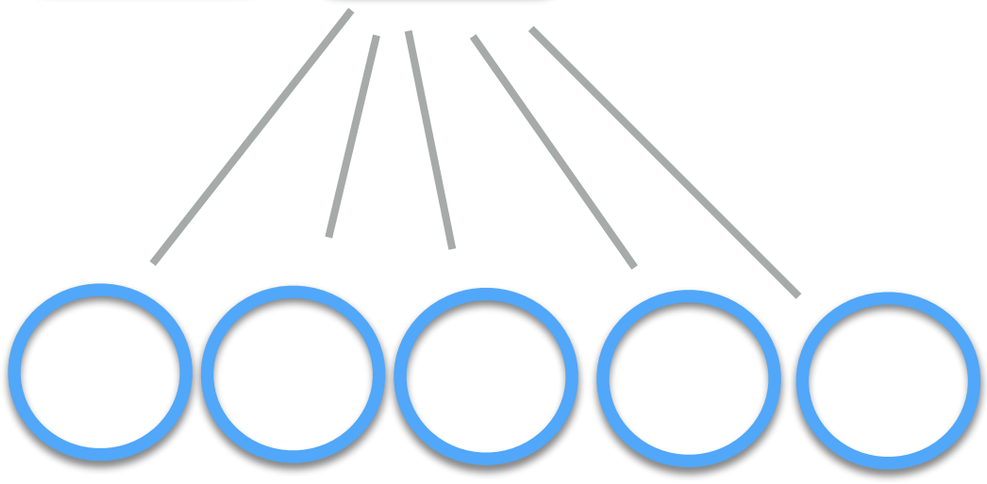
Medizinische Fakultät Heidelberg

- **4 research projects** have been defined
  - Project 01: Cancer Hallmarks
  - Project 02: Biomedical Image analysis
  - Project 03: Transcriptome analysis
  - Project 04: Lung metastasis
- Each project has **supervisors** and **tutors**
  - **supervisors**: have designed the project and know the biological background
  - **tutors**: will interact with the students (and supervisors if needed)
- Each topic has up to **5 sub-projects;** each sub-project will be worked out by **groups of 4 students**
- **Role of the master tutors:**
  - online weekly meetings with groups working on project (Wednesday 10am-1pm)
  - Online meetings will be organized using Discord
- Tutors:
  Anton Hanke (Project 01); Nicolas Peschke (Project 02); Matteo Spatuzzi (Project 03); Alvaro Mendoza (Project 04)

Medizinische Fakultät Heidelberg

Project 01 | Project 02 | Project 03 | Project 04

Sup. Tutor | Sup. Tutor | Sup. Tutor | Sup. Tutor



5 groups of 4 students

jupyter

GitHub

R Markdown

from RStudio

Projects / sub-projects

Medizinische Fakultät Heidelberg

- **Project 01**: ***Cancer Hallmarks***

  (Ashwini Sharma / Carl Herrmann)

  - Data types: gene expression / gene sets of hallmarks and metabolic pathways

- **Project 02**: ***Biomedical image analysis***

  (Christian Ritter / Karl Rohr)

  - Data types: MNIST images / cell nuclei images

- **Project 03**: ***Cancer transcriptomics***

  (Christian Heyer / Matthias Schlesner)

  - Data types: RNA-seq from TCGA

- **Project 04**: ***Aberrant expression in metastatic lung cancer***

  (Nicolas Palacio / Aurélien Dugourd / Julio Saez-Rodriguez)

  - Data types: gene expression

Timeline

	We do…	*You do…*
22/04	Presentation of the projects	
29/04	Presentation of R markdown and github	selection of projects and teams; registration
20/05		Presentation of project proposal (10 + 10 min)
mid/end July		Final presentation (15+10 minutes)

Project proposal (20/05)

Medizinische Fakultät Heidelberg

- During the project proposal presentation, you should
  - review some of the references given in the project description
  - explain what the questions / challenges are
  - describe which of these questions you want to adress in your project
  - indicate a approximate timeline
    - milestones = important steps in the analysis
    - when these milestones should be achieved

- Presentation in front of the project supervisors (video conference)
  - 10 minutes presentation
  - 10 minutes discussion / questions

- *All team members are expected to contribute!*

Final evaluation

Medizinische Fakultät Heidelberg

- Each student will be evaluated individually!
- Final grade will be
  - 30% project proposal presentation
  - 30% final presentation
  - 40% written report
- Criteria for oral presentation (for Project proposal):
  - quality of the presentation of the literature
  - understanding of the biological question
  - clarity of presentation of planned analysis and milestones
  - knowledge of the datasets
  - teamwork aspects
  - quality of the oral presentation and slides

Medizinische Fakultät Heidelberg

How to get started?

Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

Project selection / registration

Medizinische Fakultät Heidelberg

- Listen to the description of the projects / sub-projects in the online videos
- Check the description of the project on this webpage https://datascience-mobi.github.io/
- Once you have selected your team and project, **register your team** in the Google sheet https://docs.google.com/spreadsheets/d/1jZ6fissYZsaxXeWwvzSioOhV_9f4er5HUzaLABEkJEk/edit?usp=sharing
- Google Sheet for registration will open **Friday, 24/04 at 10am!**
- **"First come, first serve" rule!**
- **Selection of the projects should be done by Wednesday 29/04 10am !**

DataCamp

Medizinische Fakultät Heidelberg

- DataCamp is a fabulous ressource, hosting hundreds of online courses!
- You have the chance to get free access to this during this course: make sure to use it!
- Before starting your project, take some of the suggested courses
- These will help getting started and avoid getting stuck with simple tasks: reading a data frame, loading a file from your computer, etc...
- In case of coding difficulties, tutors will point you to the appropriate ressource.

How to (not) use the tutors?

Medizinische Fakultät Heidelberg

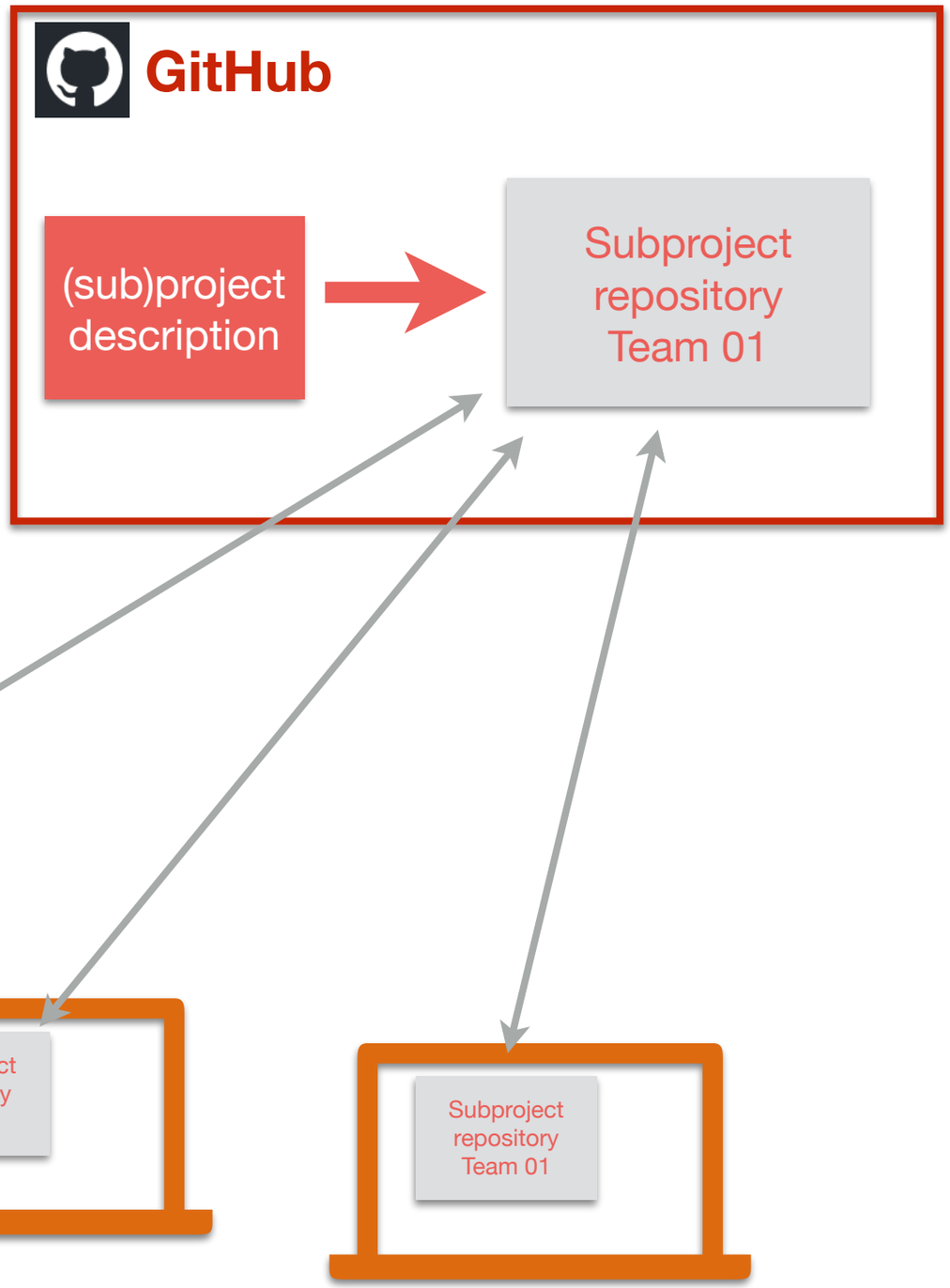
Tutors are NOT meant to answer your WhatsApp questions at midnight...

- Weekly interaction time with tutors is **limited**; make sure to use it properly
- Tutors will **NOT spend their time debugging** your code
- Prepare your meeting with the tutors, by thinking about the following points
  - what did we do since last week?
  - where did we make progress?
  - what are the problems encountered?

Organizing your work

Medizinische Fakultät Heidelberg

Website
datascience-mobi.github.io



jupyter

R Markdown
from RStudio

Medizinische Fakultät Heidelberg

Brief intro to Git(Hub)

Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

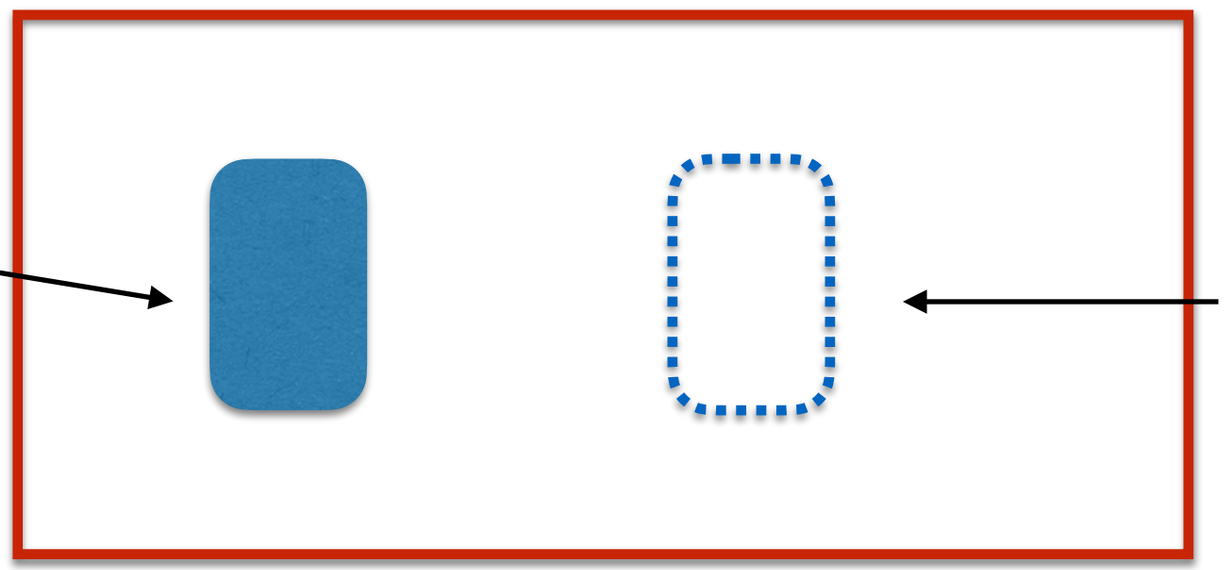
Git(Hub)

Medizinische Fakultät Heidelberg

- Git is a **version control system**:
  - allows simultaneous work of different people on the same project
  - tracks the changes (‘**commits**’) made by each member
  - helps solve the **conflicts** between various versions
- GitHub is a plateform which hosts Git projects (‘repositories’)
  - is free to use
  - required to create a (free) account
  - can be used in command line or using GUI tools (‘GitHub Desktop’)

Git repository

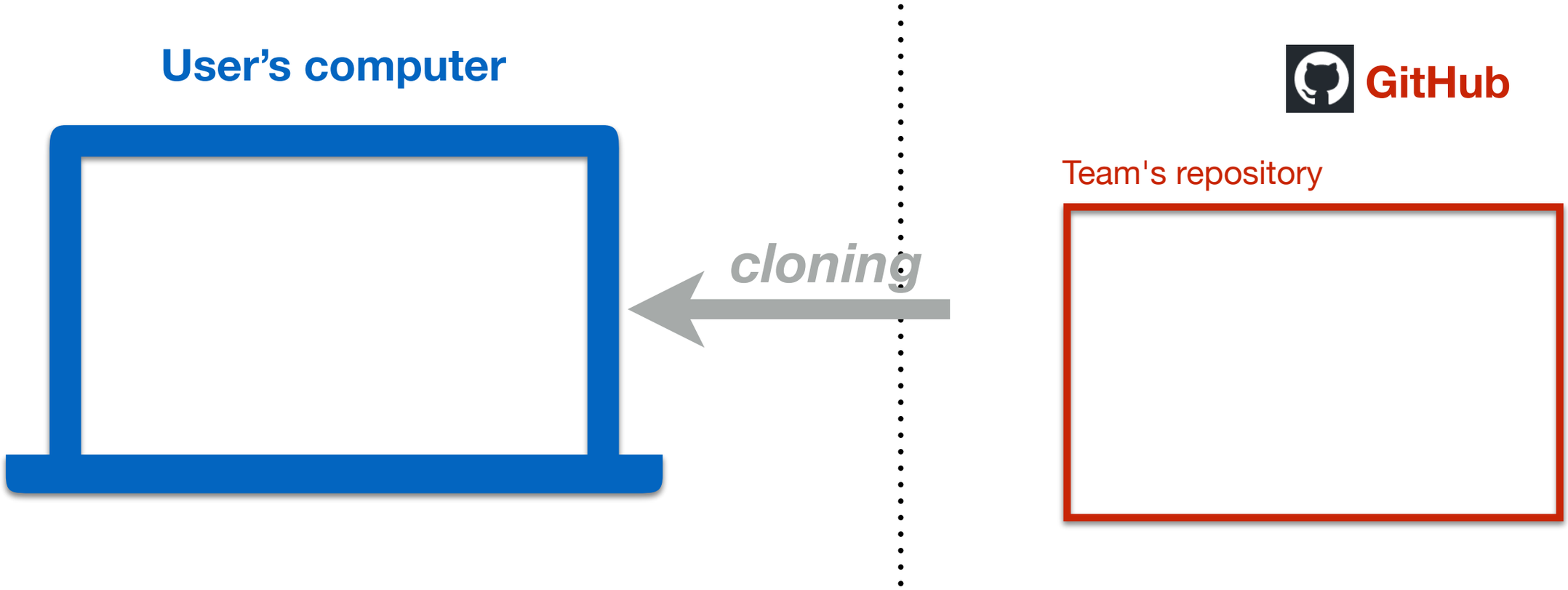
this file is registered
in the database
(‘commited’)

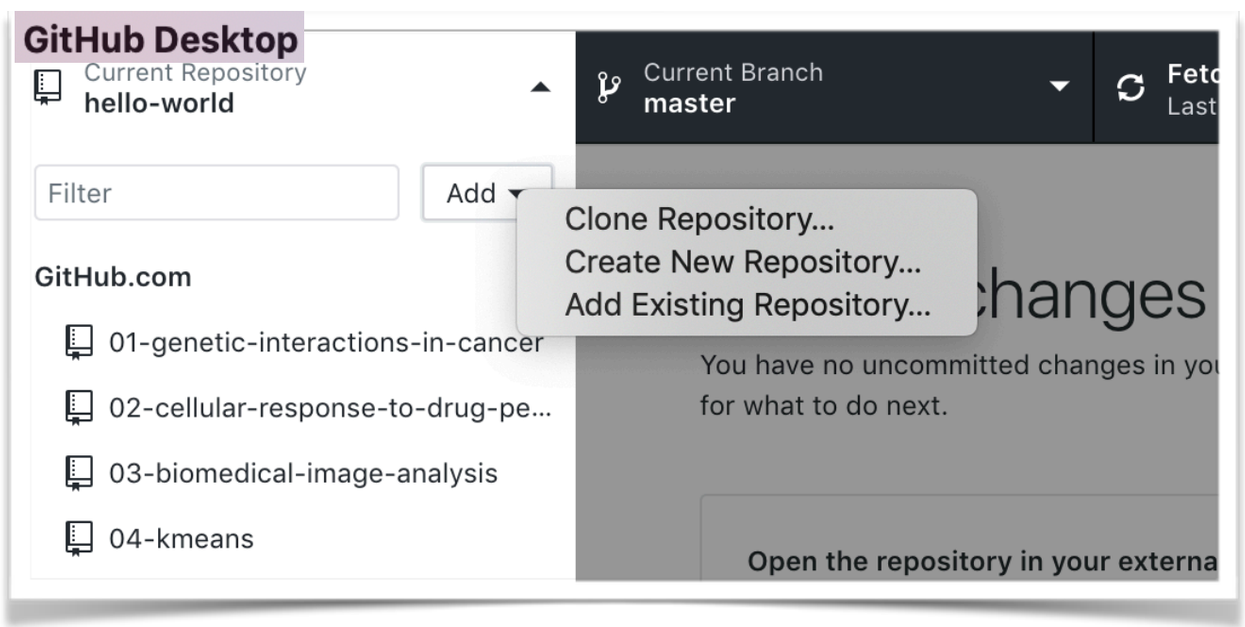


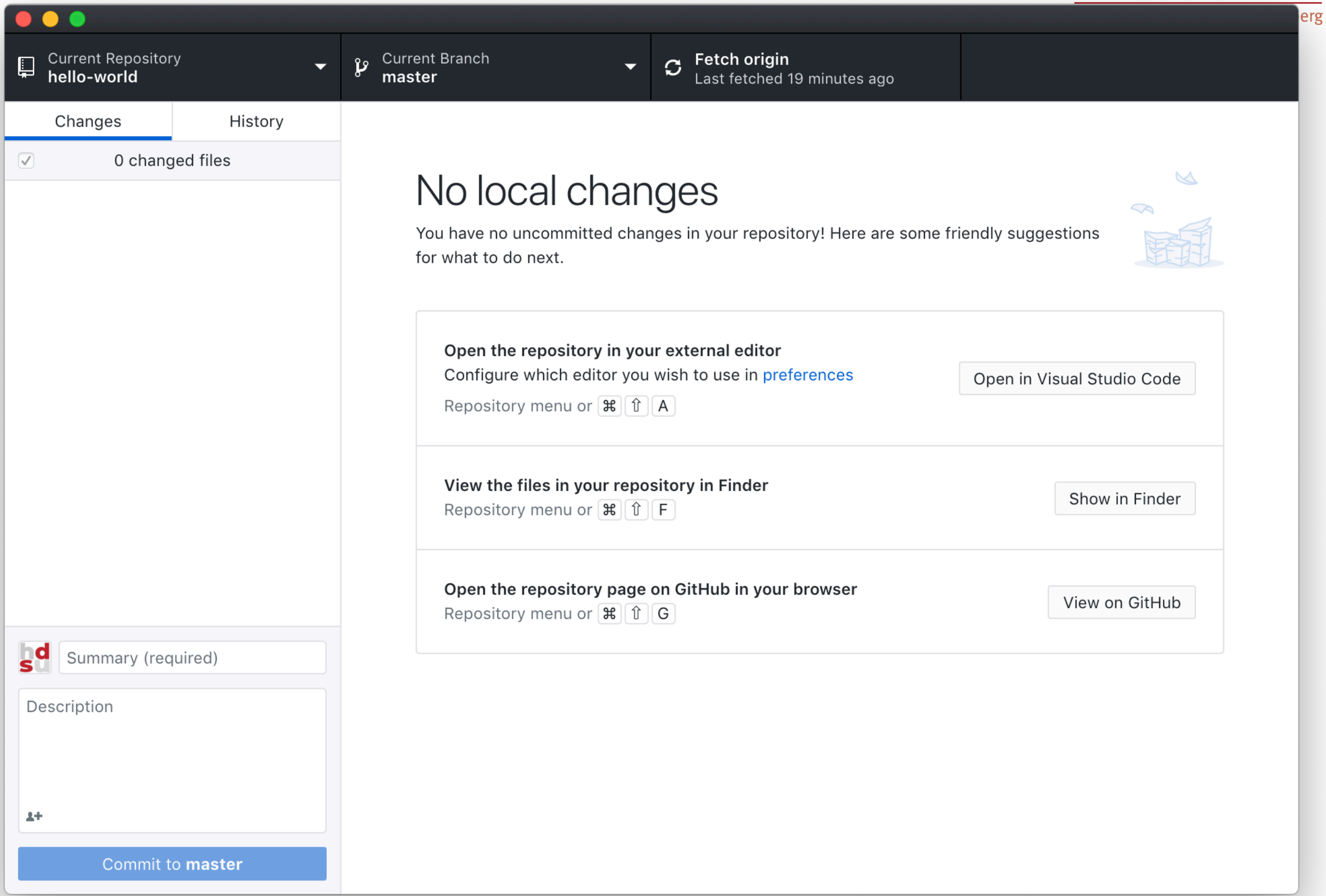
this file exists, but
has not been
registered yet
OR
the file has been modified, but
the changes have not been
registered yet

1. Cloning an existing repository

Medizinische Fakultät Heidelberg

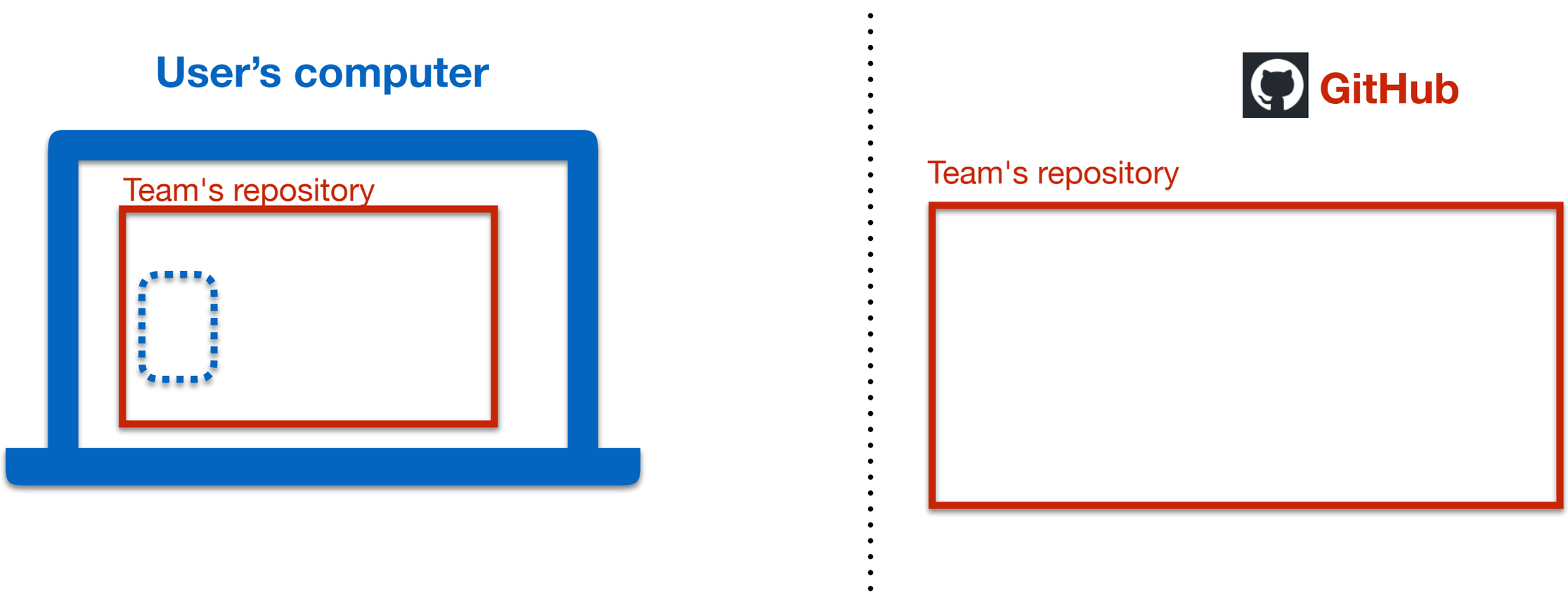




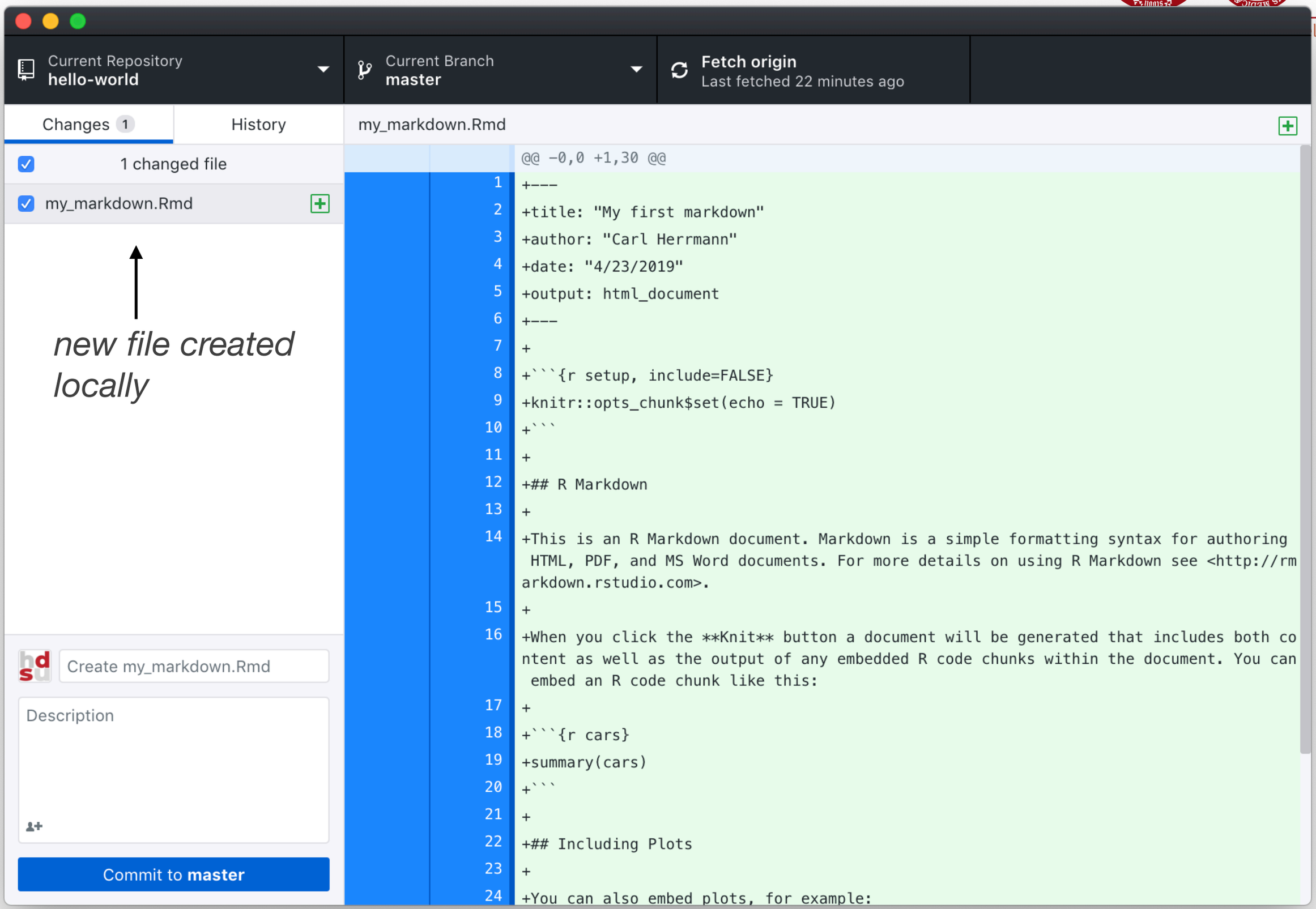

Current Repository
hello-world
Current Branch
master
Fetch origin
Last fetched 19 minutes ago
Changes
History
0 changed files
Summary (required)
Description
Commit to master
No local changes
You have no uncommitted changes in your repository! Here are some friendly suggestions for what to do next.
Open the repository in your external editor
Configure which editor you wish to use in preferences
Repository menu or ⌘ ⇧ A
Open in Visual Studio Code
View the files in your repository in Finder
Repository menu or ⌘ ⇧ F
Show in Finder
Open the repository page on GitHub in your browser
Repository menu or ⌘ ⇧ G
View on GitHub

2. creating a local file

Medizinische Fakultät Heidelberg

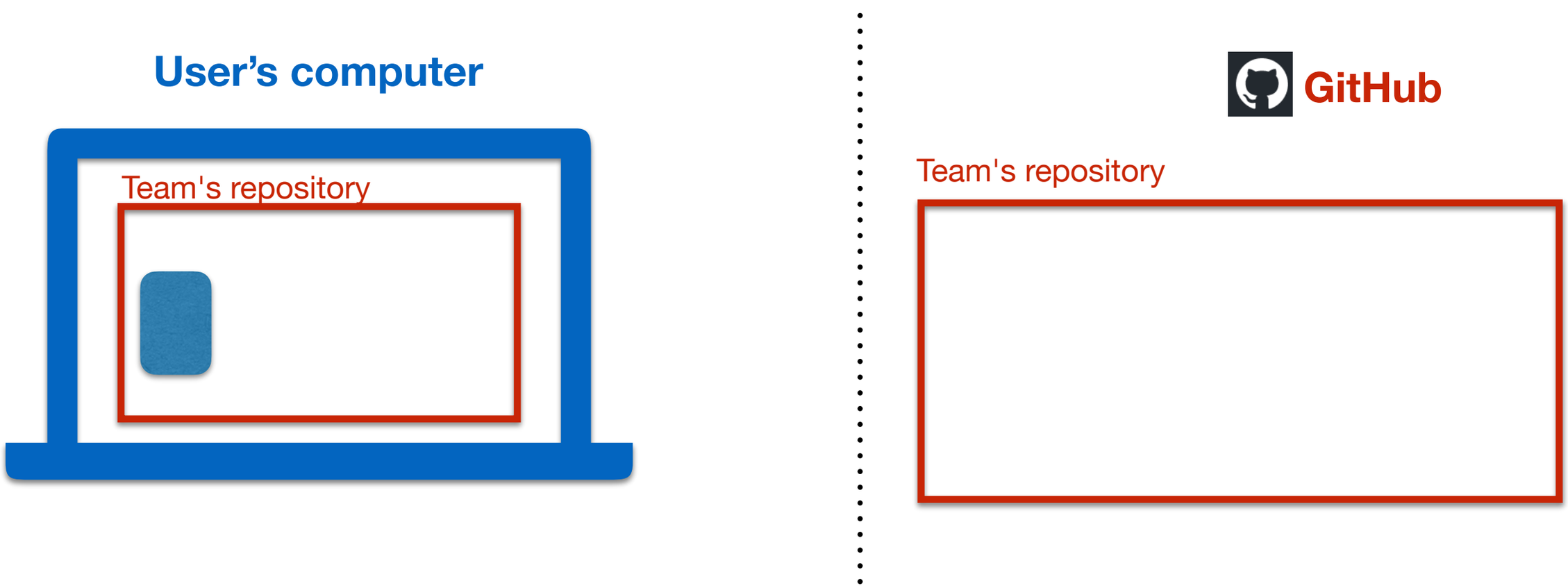


- When a new file is added / modified in the local folder, it is not yet registered in the git database!
- it first needs to be **commited**

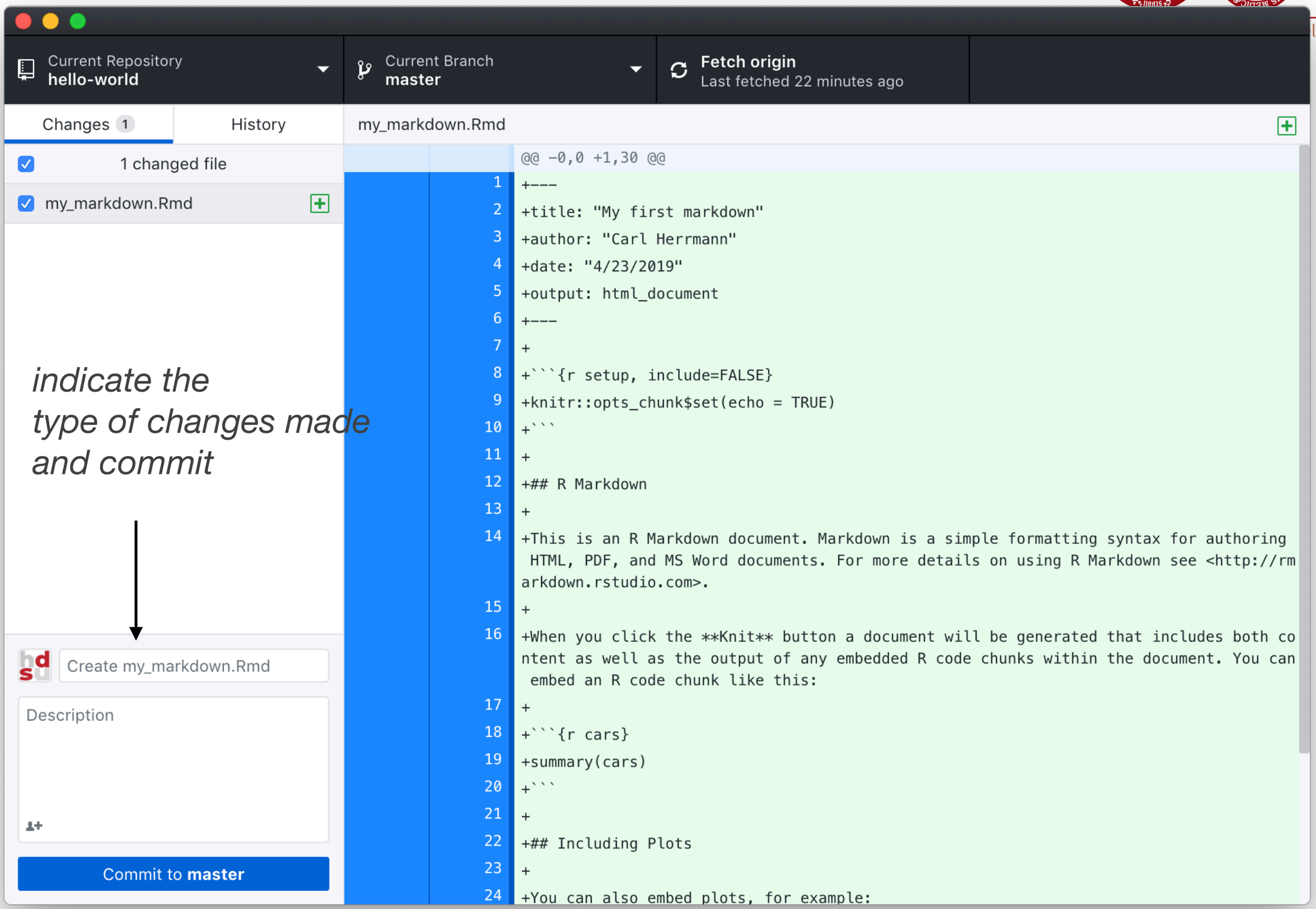

Current Repository
hello-world
Current Branch
master
Fetch origin
Last fetched 22 minutes ago
Changes 1
History
my_markdown.Rmd
1 changed file
my_markdown.Rmd
new file created locally
Create my_markdown.Rmd
Description
Commit to master
@@ -0,0 +1,30 @@
+---
+title: "My first markdown"
+author: "Carl Herrmann"
+date: "4/23/2019"
+output: html_document
+---
+
+```{r setup, include=FALSE}
+knitr::opts_chunk$set(echo = TRUE)
+```
+
+## R Markdown
+
+This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
+
+When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
+
+```{r cars}
+summary(cars)
+```
+
+## Including Plots
+
+You can also embed plots, for example:

2. Adding a file

Medizinische Fakultät Heidelberg

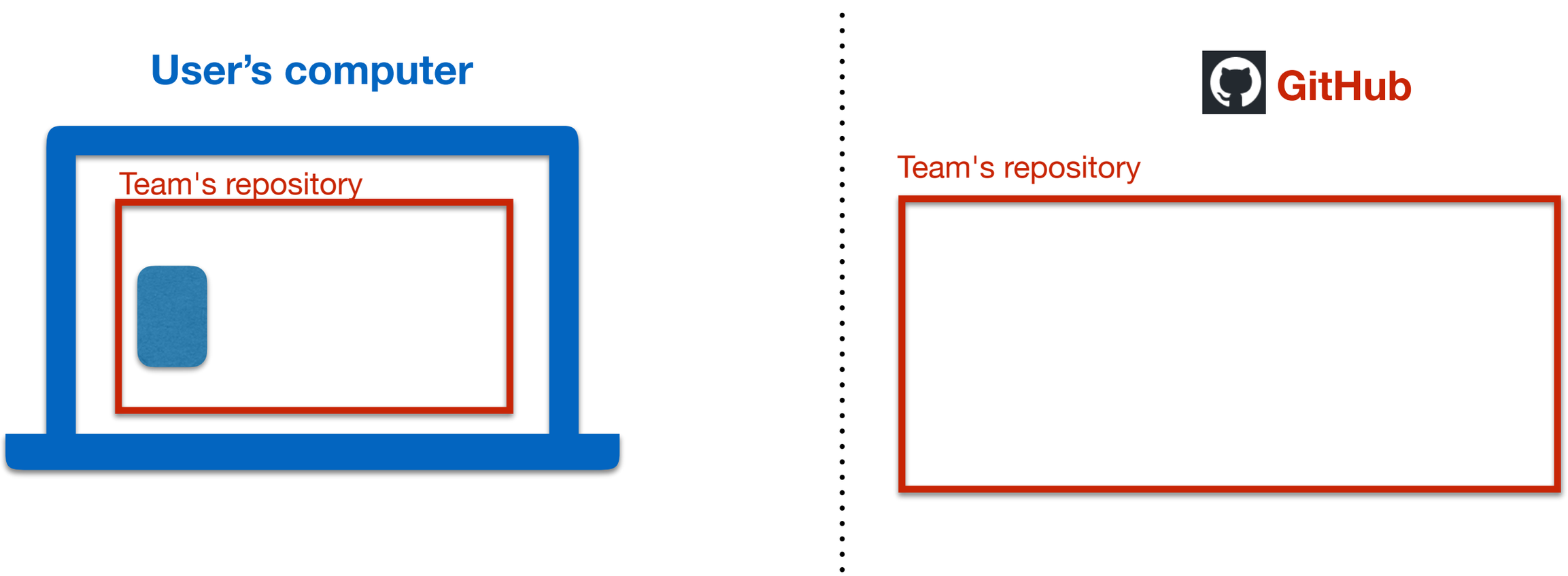


- When a new file is added / modified in the local folder, it is not yet registered in the git database!
- it first needs to be **commited**


Current Repository
hello-world
Current Branch
master
Fetch origin
Last fetched 22 minutes ago
Changes 1
History
my_markdown.Rmd
1 changed file
my_markdown.Rmd
indicate the
type of changes made
and commit
Create my_markdown.Rmd
Description
Commit to master
@@ -0,0 +1,30 @@
+---
+title: "My first markdown"
+author: "Carl Herrmann"
+date: "4/23/2019"
+output: html_document
+---
+
+```{r setup, include=FALSE}
+knitr::opts_chunk$set(echo = TRUE)
+```
+
+## R Markdown
+
+This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
+
+When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
+
+```{r cars}
+summary(cars)
+```
+
+## Including Plots
+
+You can also embed plots, for example:

2. Adding a file

Medizinische Fakultät Heidelberg



- the file is now commited to the local git repository
- it needs to be pushed to the remote repository on GitHub

Medizinische Fakultät Heidelberg

Current Repository
hello-world

Current Branch
master

Push origin
Last fetched 30 minutes ago
1

Changes | History

0 changed files

Summary (required)

Description

Commit to **master**

Committed just now
Create my_markdown.Rmd
Undo

No local changes

You have no uncommitted changes in your repository! Here are some friendly suggestions for what to do next.

Push 1 commit to the origin remote
You have one local commit waiting to be pushed to GitHub
Always available in the toolbar when there are local commits waiting to be pushed or ⌘ P

Push origin

Open the repository in your external editor
Configure which editor you wish to use in preferences
Repository menu or ⌘ ⇧ A

Open in Visual Studio Code

View the files in your repository in Finder
Repository menu or ⌘ ⇧ F

Show in Finder

Open the repository page on GitHub in your browser
Repository menu or ⌘ ⇧ G

View on GitHub

Medizinische Fakultät Heidelberg

test repository

Edit

Manage topics

2 commits | 1 branch | 0 releases | 1 contributor

Branch: master | New pull request | Create new file | Upload files | Find File | Clone or download

carlherrmann Create my_markdown.Rmd — Latest commit b2b0bdf 2 minutes ago

README.md	Initial commit	37 minutes ago
my_markdown.Rmd	Create my_markdown.Rmd	2 minutes ago

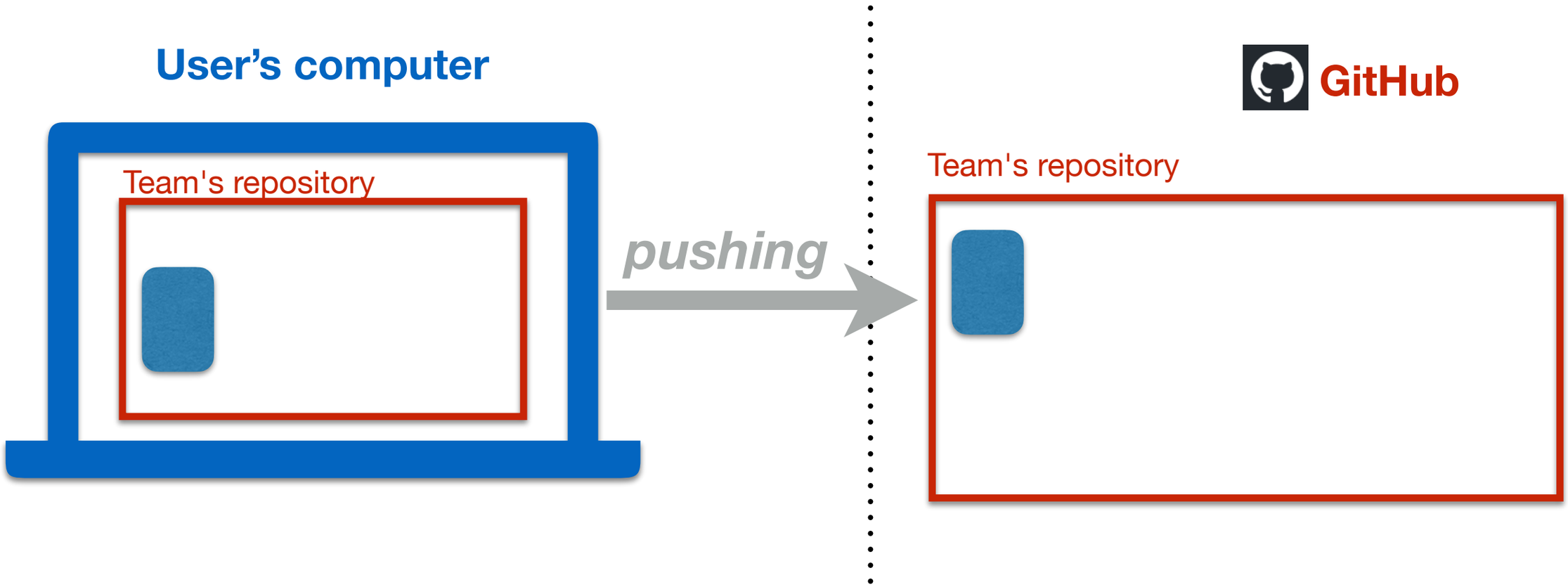
README.md

hello-world

test repository

2. Adding a file

Medizinische Fakultät Heidelberg

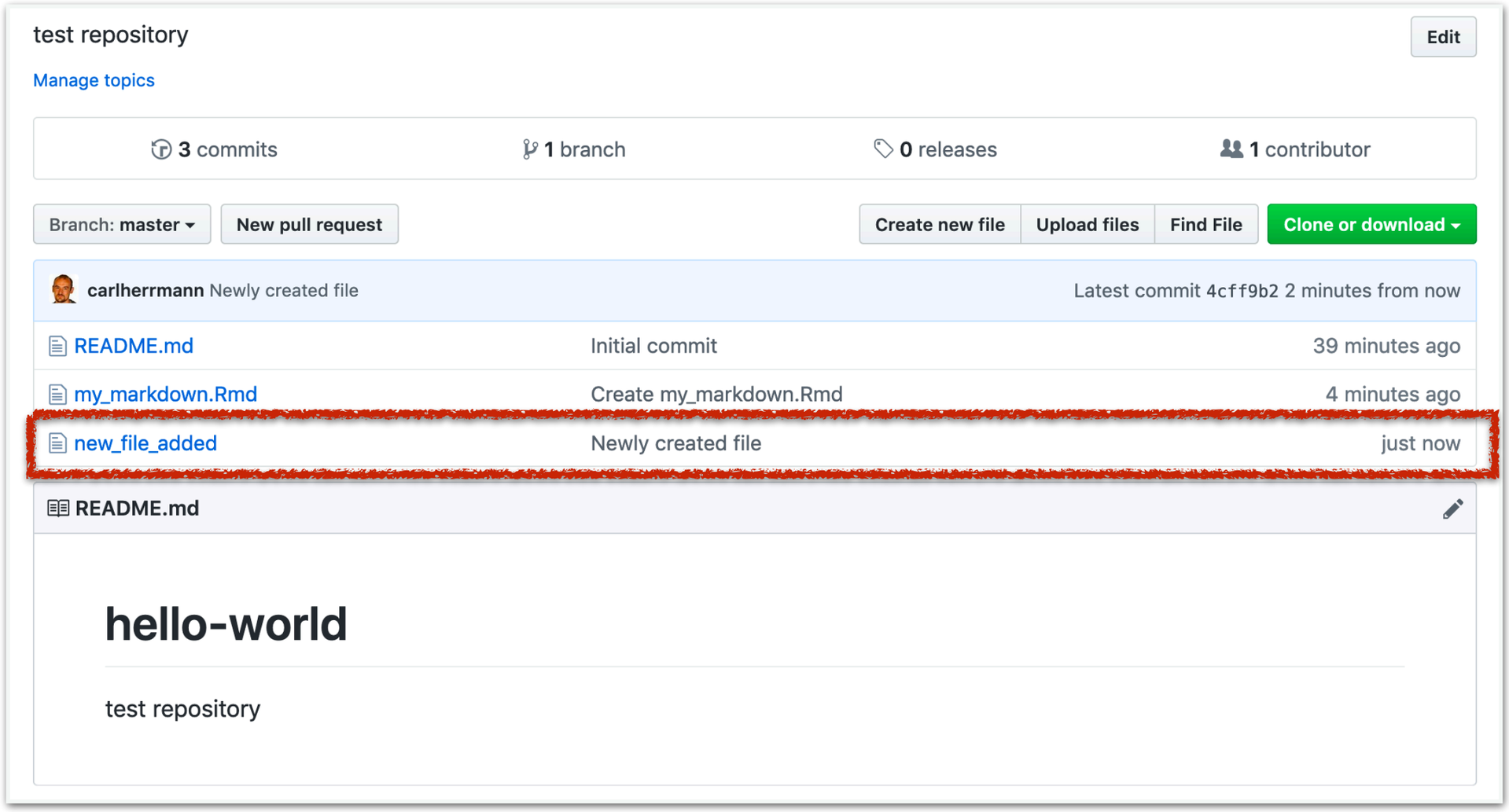


- the file is no commited to the local git repository
- it needs to be pushed to the remote repository on GitHub

3. Pulling from the remote repository

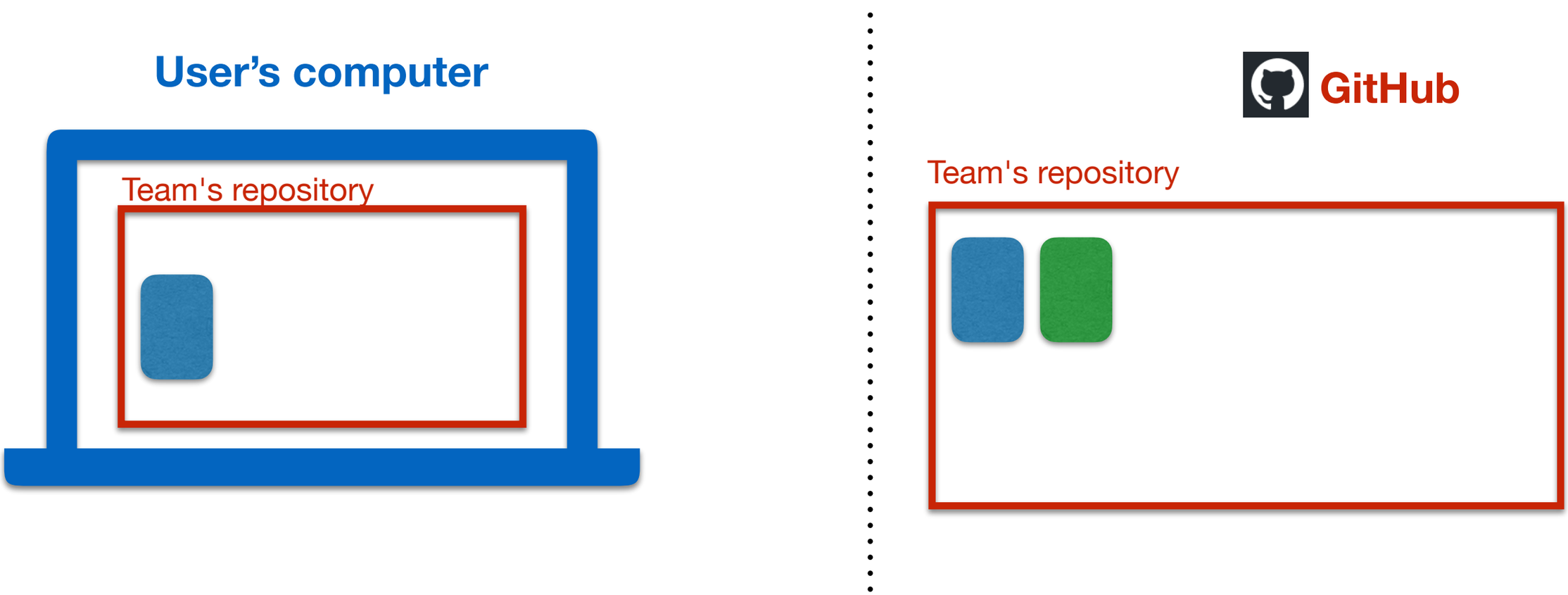
Medizinische Fakultät Heidelberg

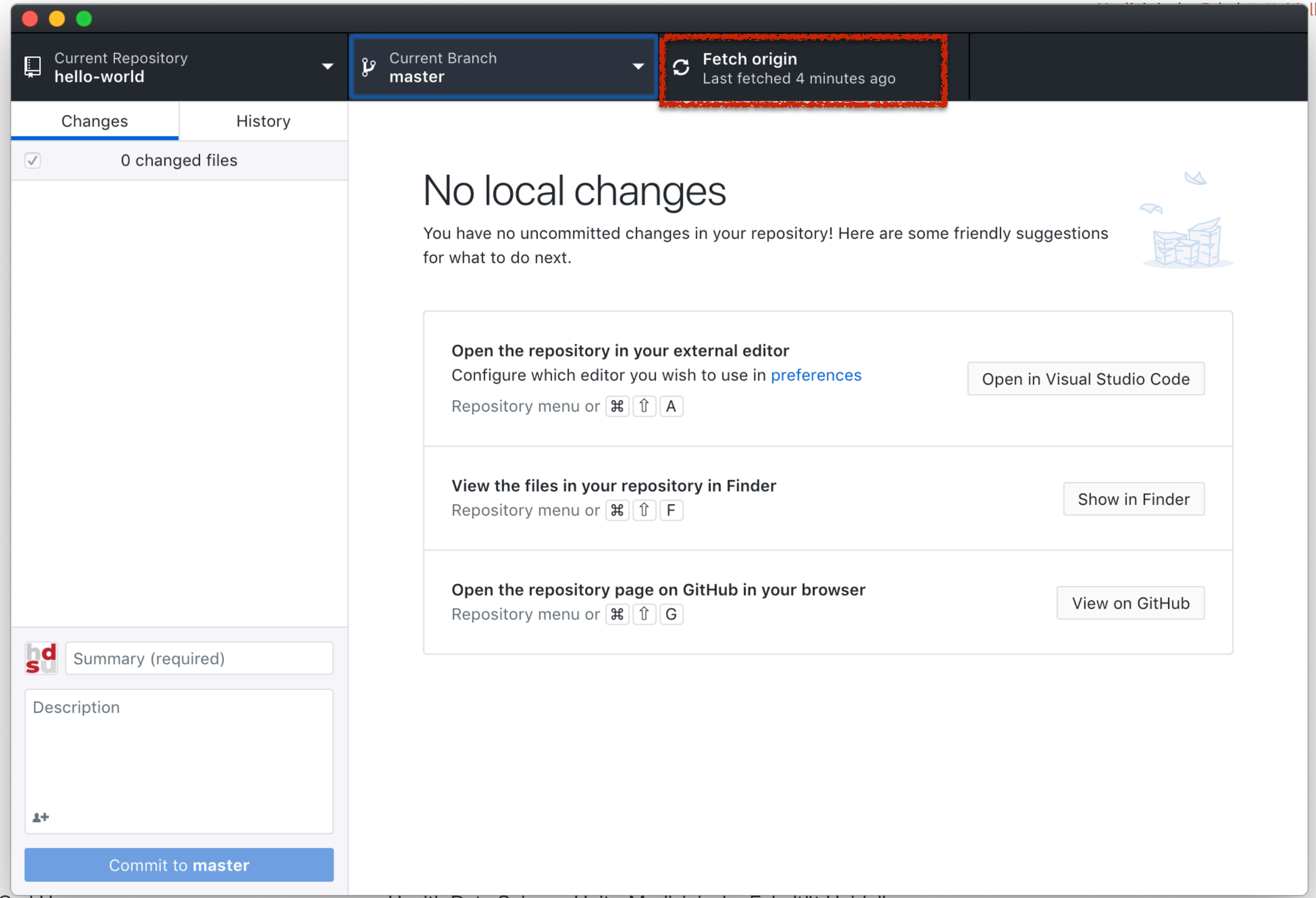
- Someone (probably one of your team mates) has added a new file into the remote repository
- It is not yet in your local repository and need to be **pulled**

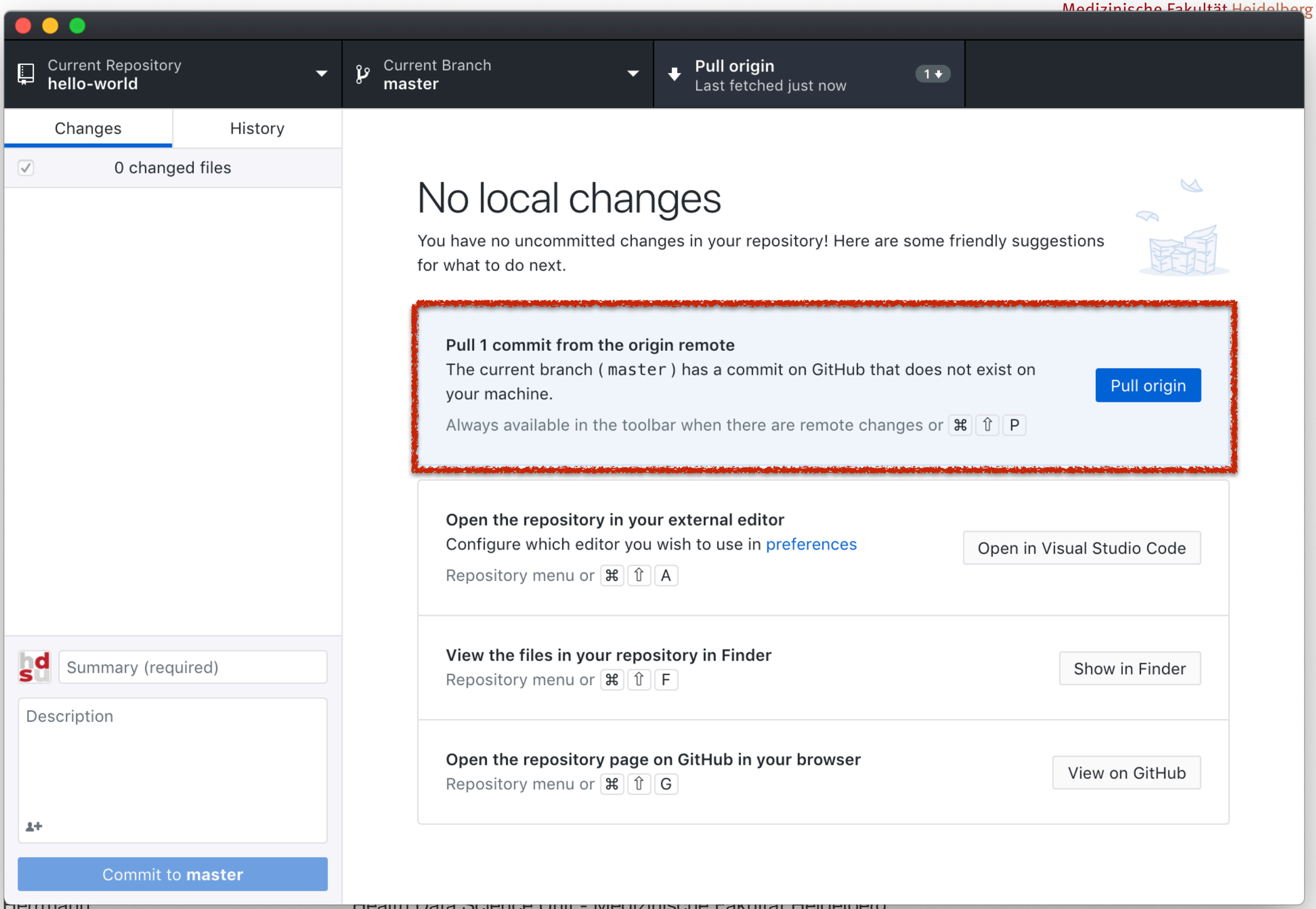


3. Pulling from the remote repository

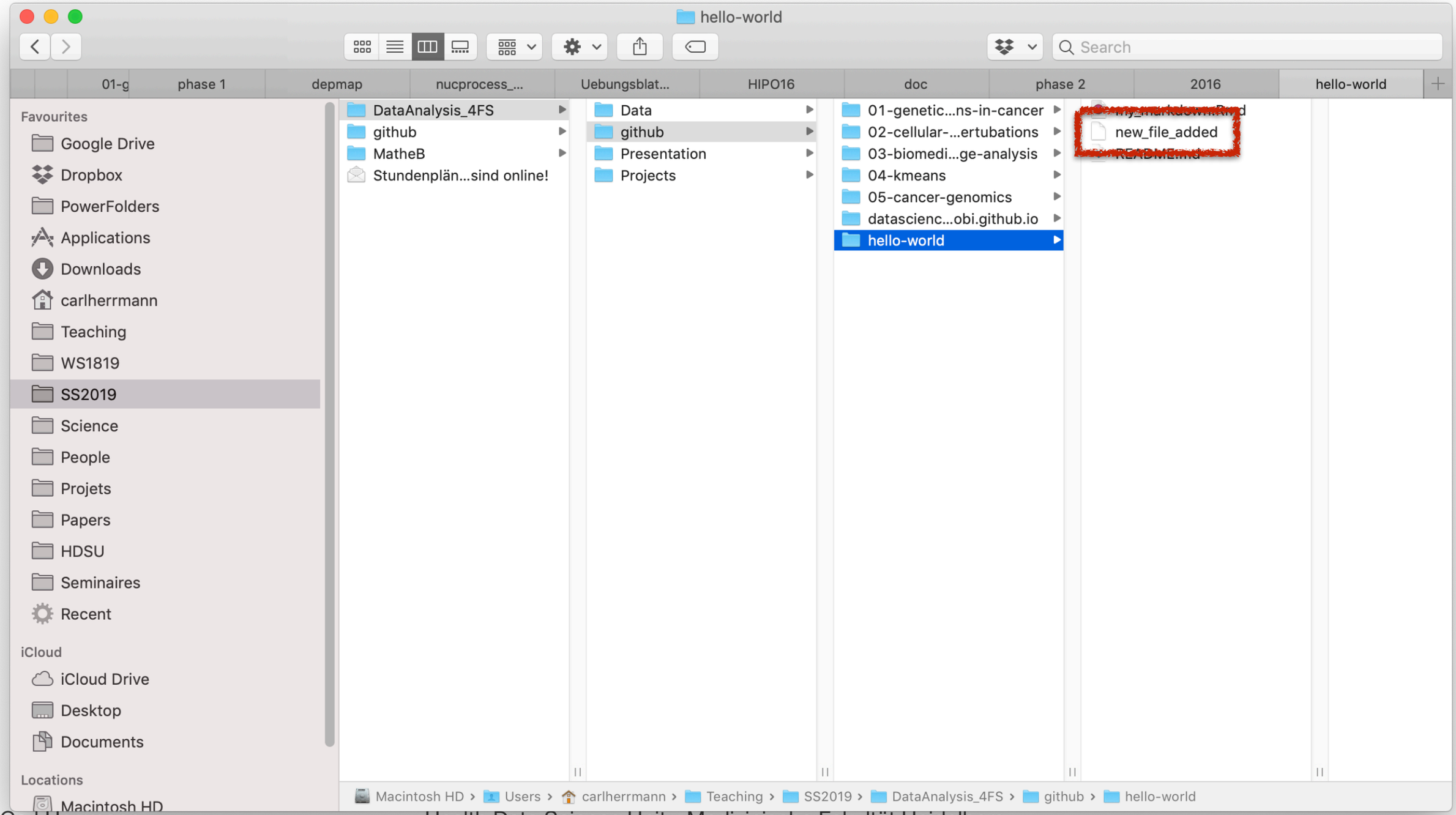
Medizinische Fakultät Heidelberg




Current Repository
hello-world
Current Branch
master
Fetch origin
Last fetched 4 minutes ago
Changes
History
0 changed files
Summary (required)
Description
Commit to master
No local changes
You have no uncommitted changes in your repository! Here are some friendly suggestions for what to do next.
Open the repository in your external editor
Configure which editor you wish to use in preferences
Repository menu or ⌘ ⇧ A
Open in Visual Studio Code
View the files in your repository in Finder
Repository menu or ⌘ ⇧ F
Show in Finder
Open the repository page on GitHub in your browser
Repository menu or ⌘ ⇧ G
View on GitHub

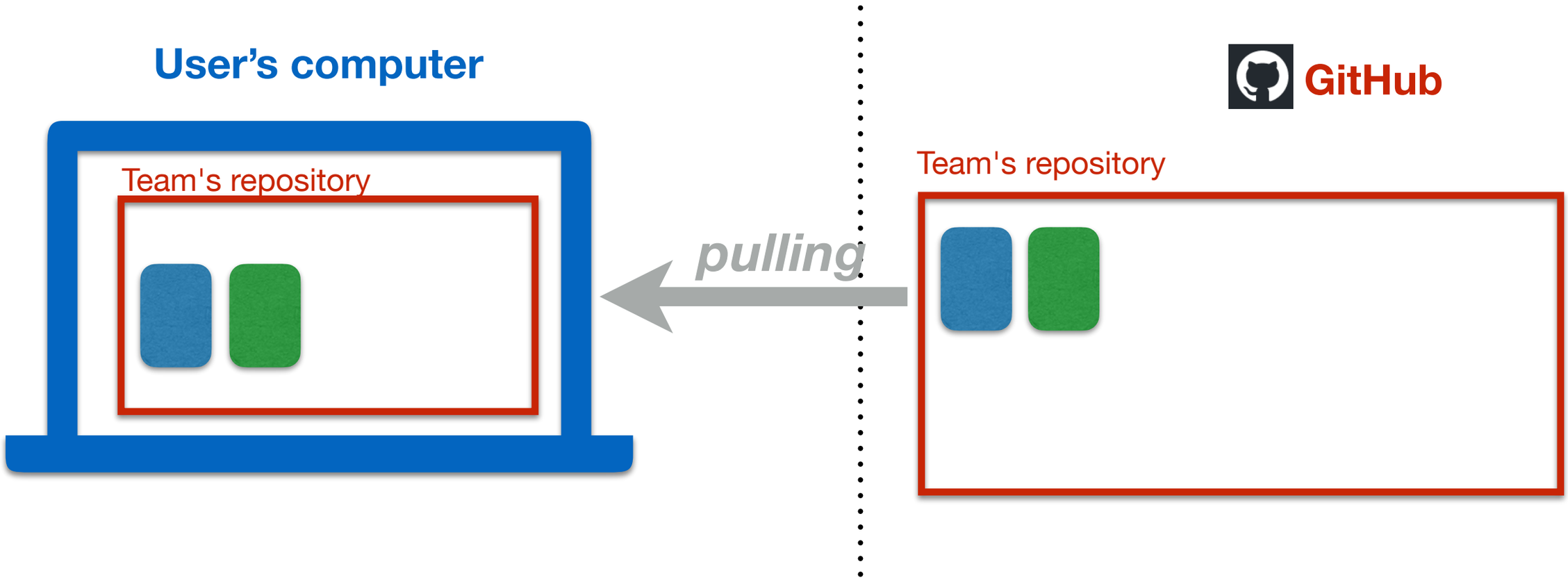

Current Repository
hello-world
Current Branch
master
Pull origin
Last fetched just now
1
Changes
History
0 changed files
Summary (required)
Description
Commit to master
No local changes
You have no uncommitted changes in your repository! Here are some friendly suggestions for what to do next.
Pull 1 commit from the origin remote
The current branch (master) has a commit on GitHub that does not exist on your machine.
Pull origin
Always available in the toolbar when there are remote changes or ⌘ ⇧ P
Open the repository in your external editor
Configure which editor you wish to use in preferences
Repository menu or ⌘ ⇧ A
Open in Visual Studio Code
View the files in your repository in Finder
Repository menu or ⌘ ⇧ F
Show in Finder
Open the repository page on GitHub in your browser
Repository menu or ⌘ ⇧ G
View on GitHub

- Once the remote repository is pulled, the new file(s) are available locally



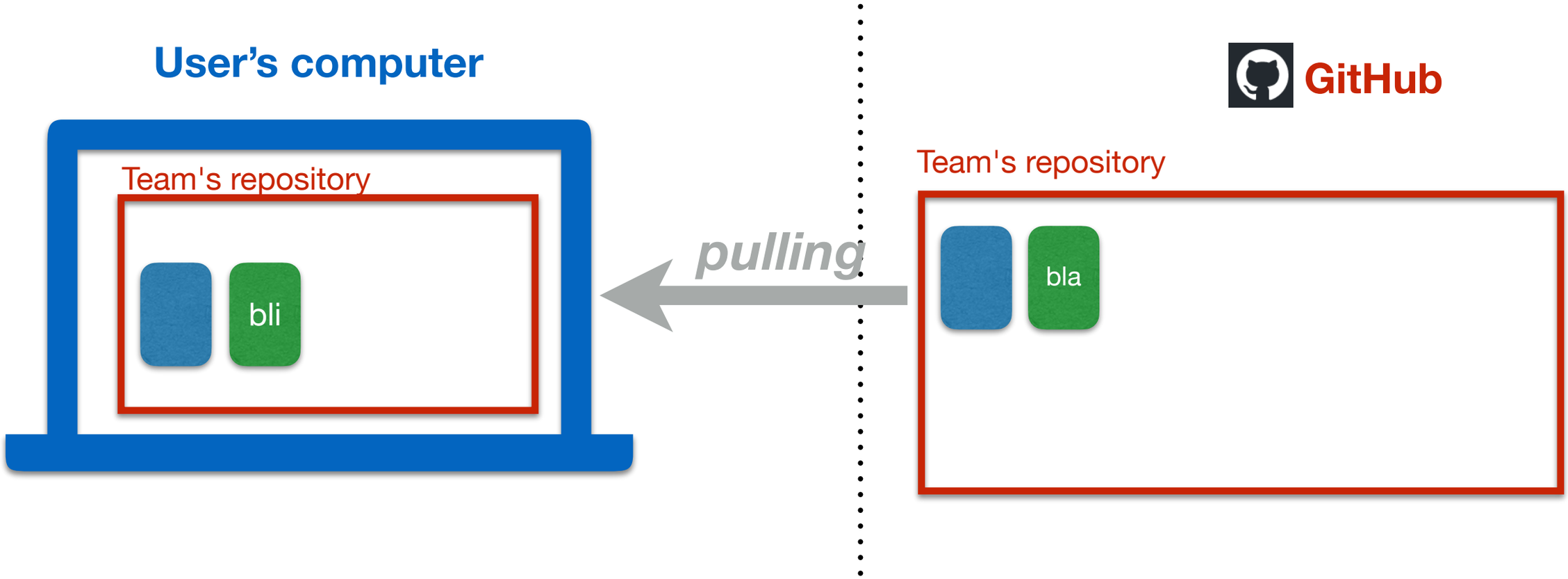
3. Pulling from the remote repository

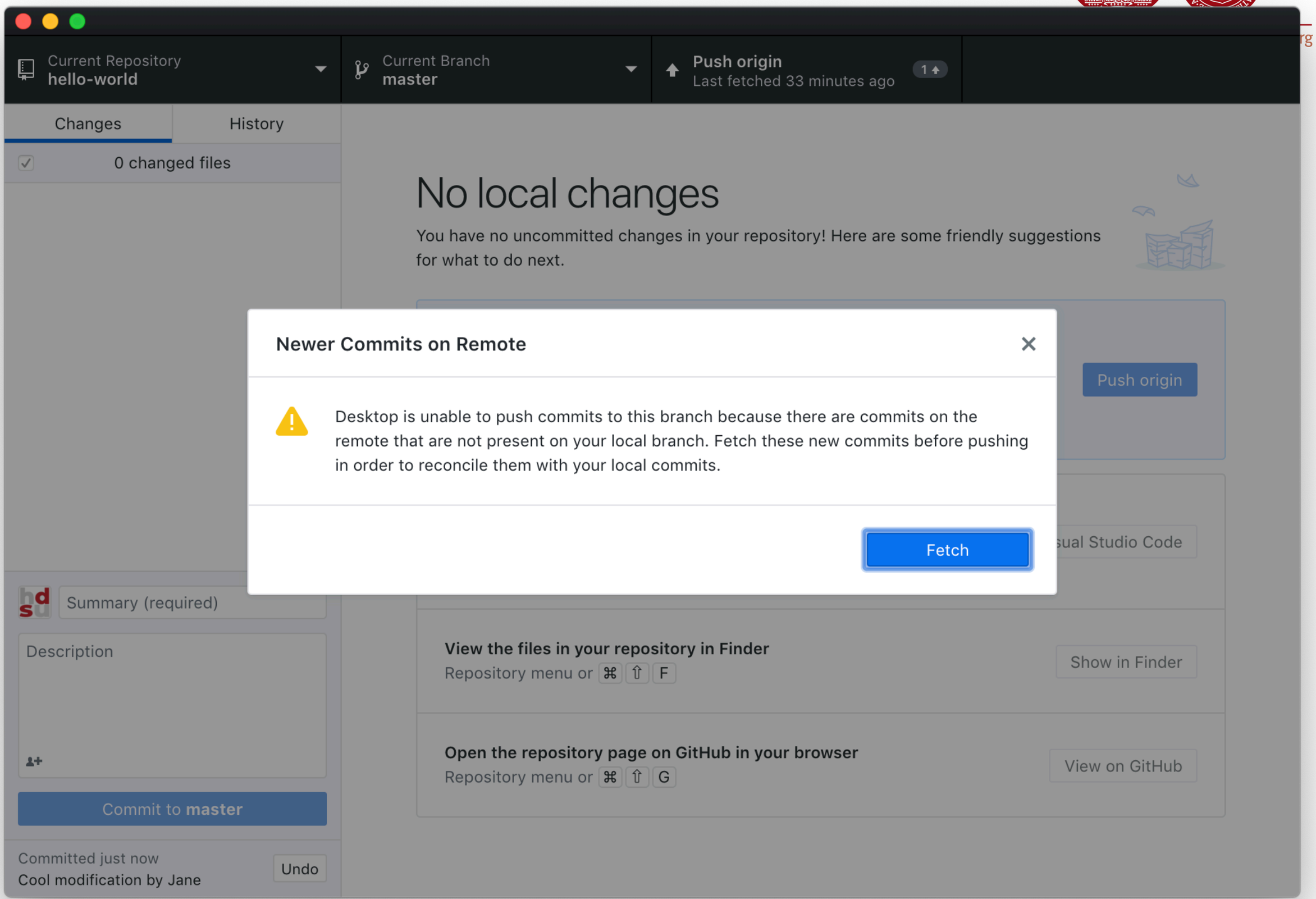
Medizinische Fakultät Heidelberg

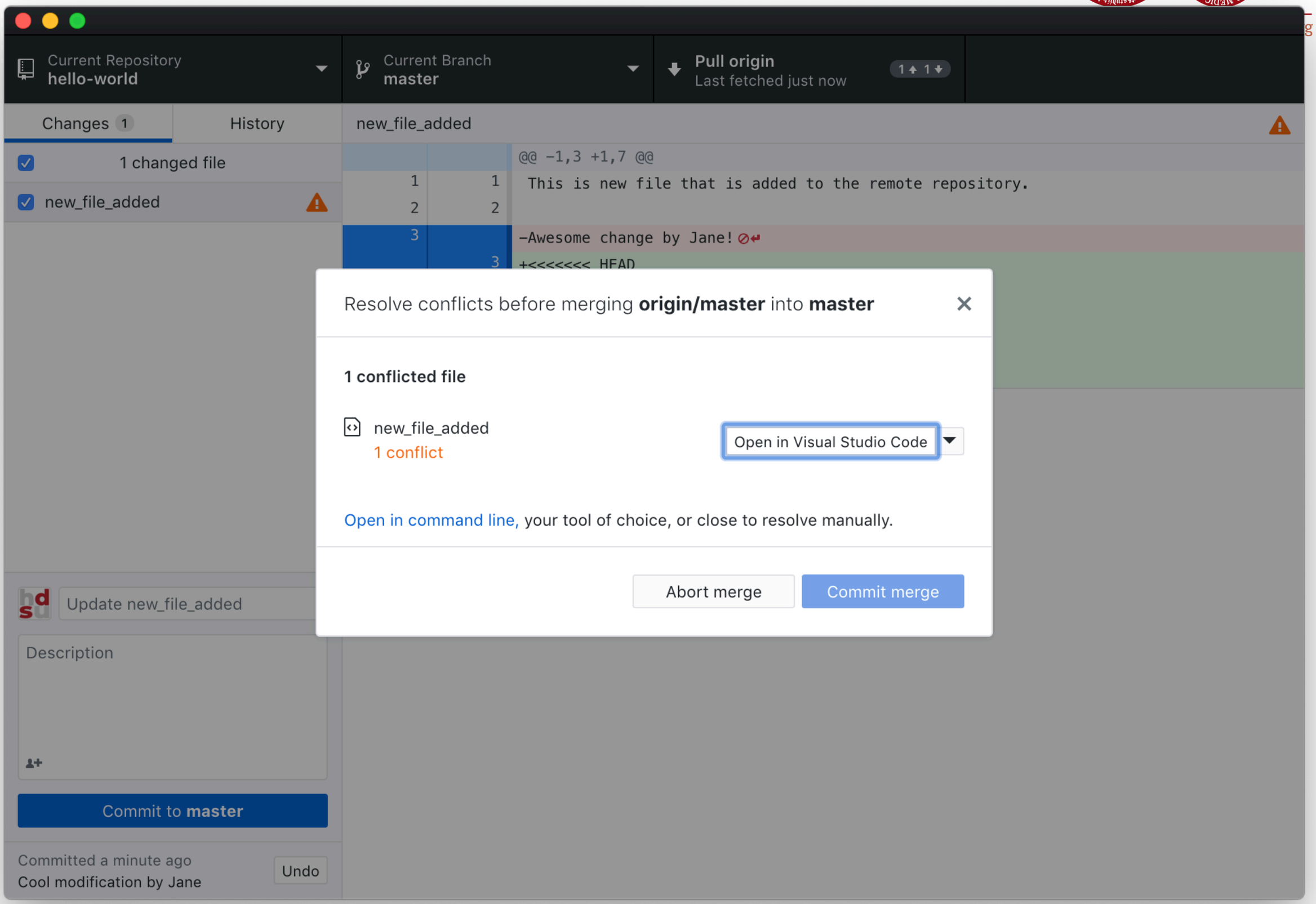


4. conflicting changes

Medizinische Fakultät Heidelberg




Current Repository
hello-world
Current Branch
master
Push origin
Last fetched 33 minutes ago
1
Changes
History
0 changed files
No local changes
You have no uncommitted changes in your repository! Here are some friendly suggestions for what to do next.
Push origin
Newer Commits on Remote
Desktop is unable to push commits to this branch because there are commits on the remote that are not present on your local branch. Fetch these new commits before pushing in order to reconcile them with your local commits.
Fetch
sual Studio Code
Summary (required)
Description
Commit to master
View the files in your repository in Finder
Repository menu or ⌘ ⇧ F
Show in Finder
Open the repository page on GitHub in your browser
Repository menu or ⌘ ⇧ G
View on GitHub
Committed just now
Cool modification by Jane
Undo


Current Repository
hello-world
Current Branch
master
Pull origin
Last fetched just now
1 ↑ 1 ↓
Changes 1
History
new_file_added
1 changed file
new_file_added
@@ -1,3 +1,7 @@
1 1 This is new file that is added to the remote repository.
2 2
3 -Awesome change by Jane!
3 +<<<<<<< HEAD
Resolve conflicts before merging origin/master into master
1 conflicted file
new_file_added
1 conflict
Open in Visual Studio Code
Open in command line, your tool of choice, or close to resolve manually.
Abort merge
Commit merge
Update new_file_added
Description
Commit to master
Committed a minute ago
Cool modification by Jane
Undo

4. conflicting changes

- Conflicting changes can be resolved with a text editor
- options depend on which editor is used

```
You, a few seconds ago | 2 authors (You and others)
This is new file that is added to the remote repository.

Accept Current Change | Accept Incoming Change | Accept Both Changes | Compare Changes
<<<<<<< HEAD (Current Change)
Awesome change by Jane!
=======
This is a great new modification by Joe!
>>>>>>> af5e9c9981b21a39ed11d09f468bea576d669191 (Incoming Change)
```

local change

changes in the remote file

To do

- Create your own personal GitHub account

- Register your Github user name into the Google Sheet
  - all team members will be added to the corresponding GitHub repo
    - Project 03 - Team 02 → **`project-03-group-02`**

Medizinische Fakultät Heidelberg

(R)markdown

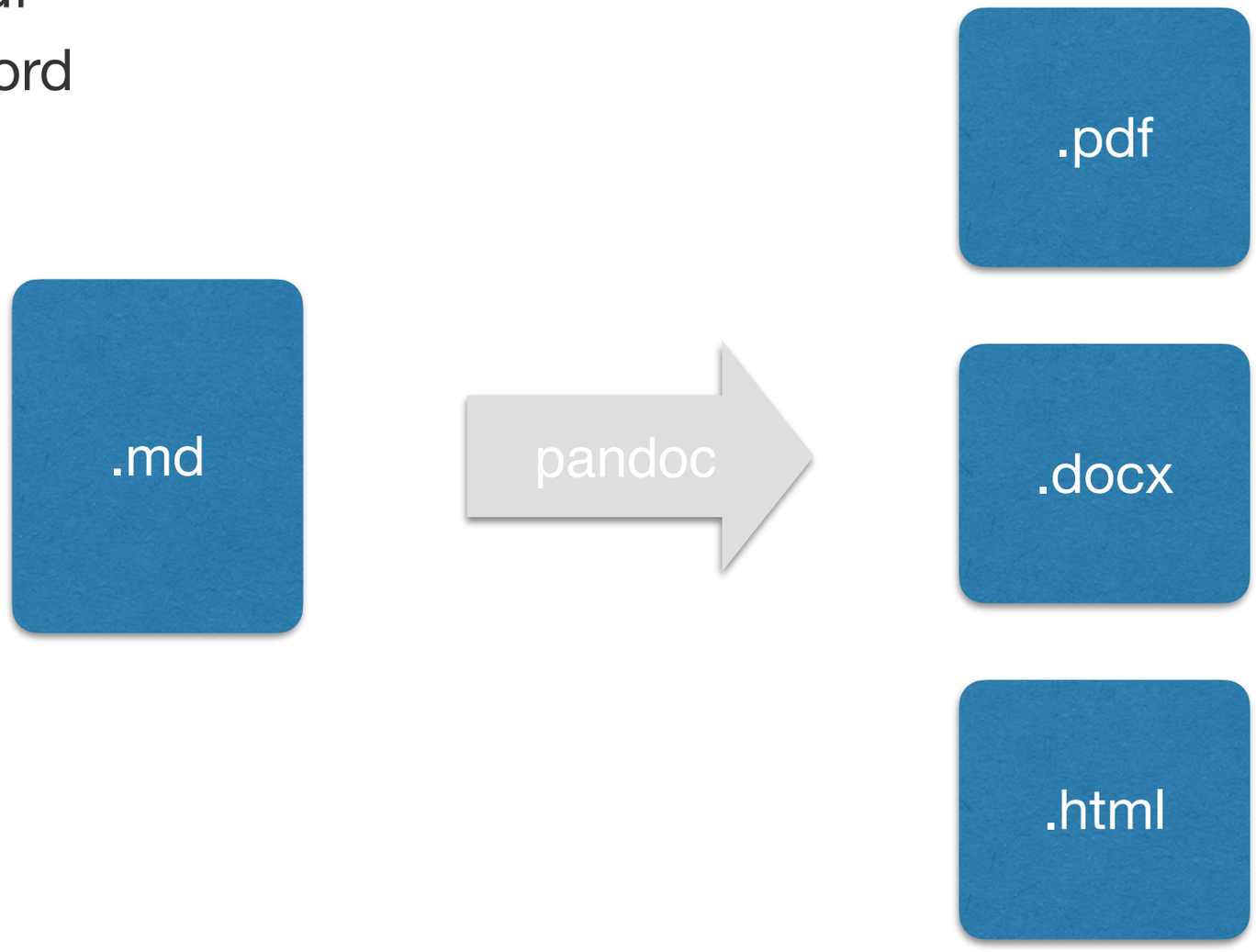
Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

Markdown

Medizinische Fakultät Heidelberg

- Markdown is a way to format plain text with a simple text editor
- Markdown documents can be converted with a **renderer** into
  - html
  - pdf
  - word



Rendering markdown

Medizinische Fakultät Heidelberg

markdown

```
# My document

## this is a header

In the text we can *highlight* or put in **bold**.

## making lists

We can make **numbered lists**:

1. first item
2. second item

or unordered lists

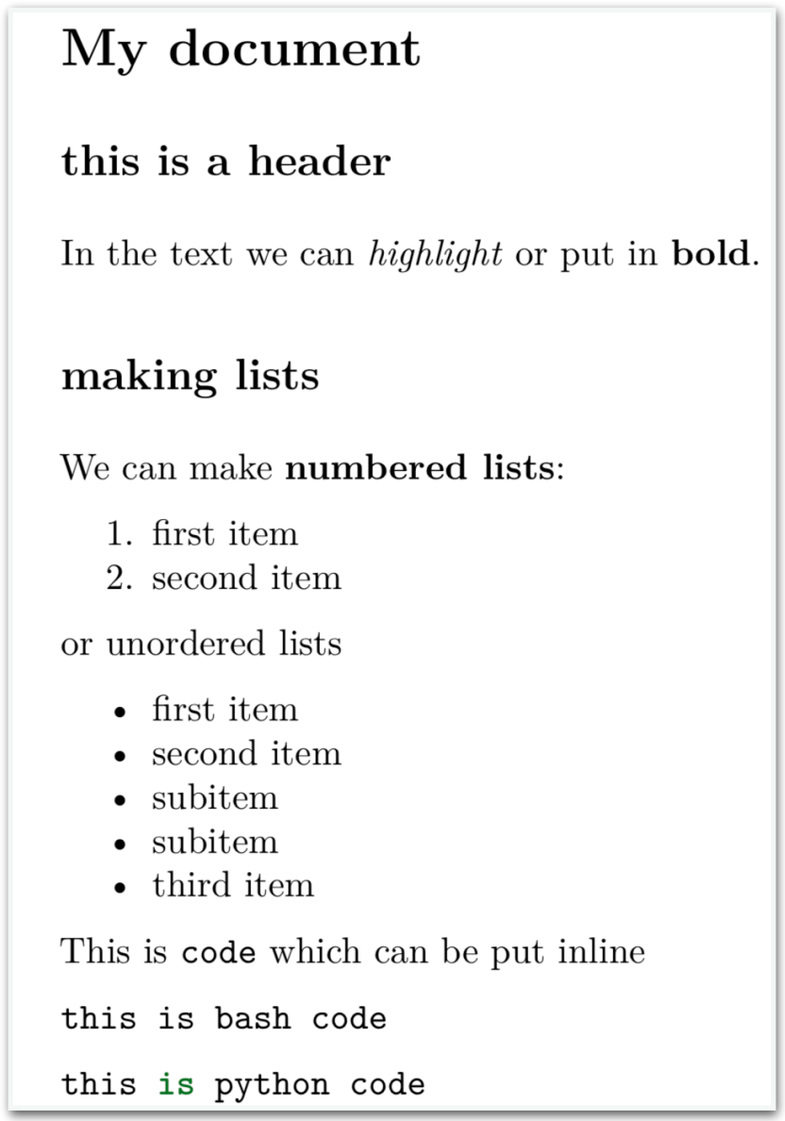
* first item
* second item
  + subitem
  + subitem
* third item

This is `code` which can be put inline

```bash
this is bash code
```

```python
this is python code
```
```

pdf



My document

this is a header

In the text we can *highlight* or put in **bold**.

making lists

We can make **numbered lists**:

1. first item
2. second item

or unordered lists

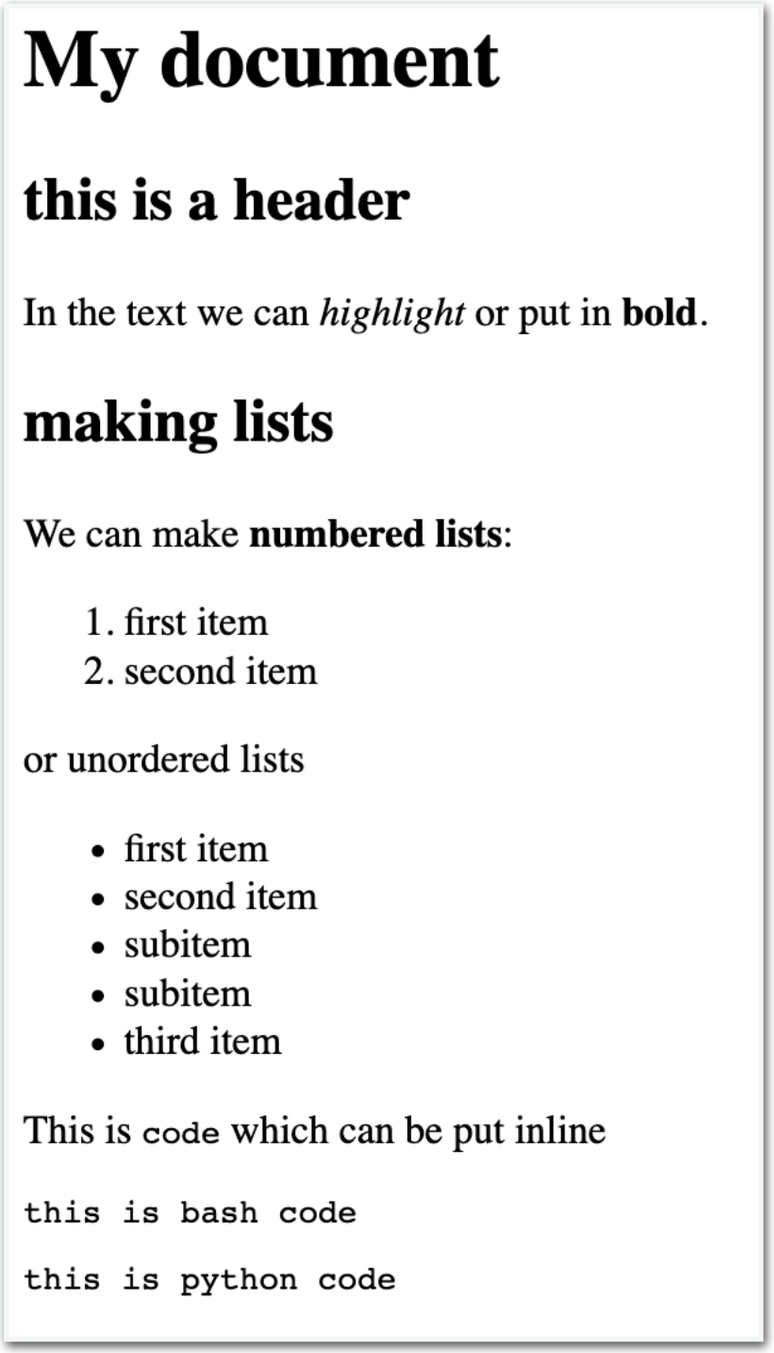
- first item
- second item
- subitem
- subitem
- third item

This is `code` which can be put inline

```
this is bash code
```

```
this is python code
```

html



My document

this is a header

In the text we can *highlight* or put in **bold**.

making lists

We can make **numbered lists**:

1. first item
2. second item

or unordered lists

- first item
- second item
- subitem
- subitem
- third item

This is `code` which can be put inline

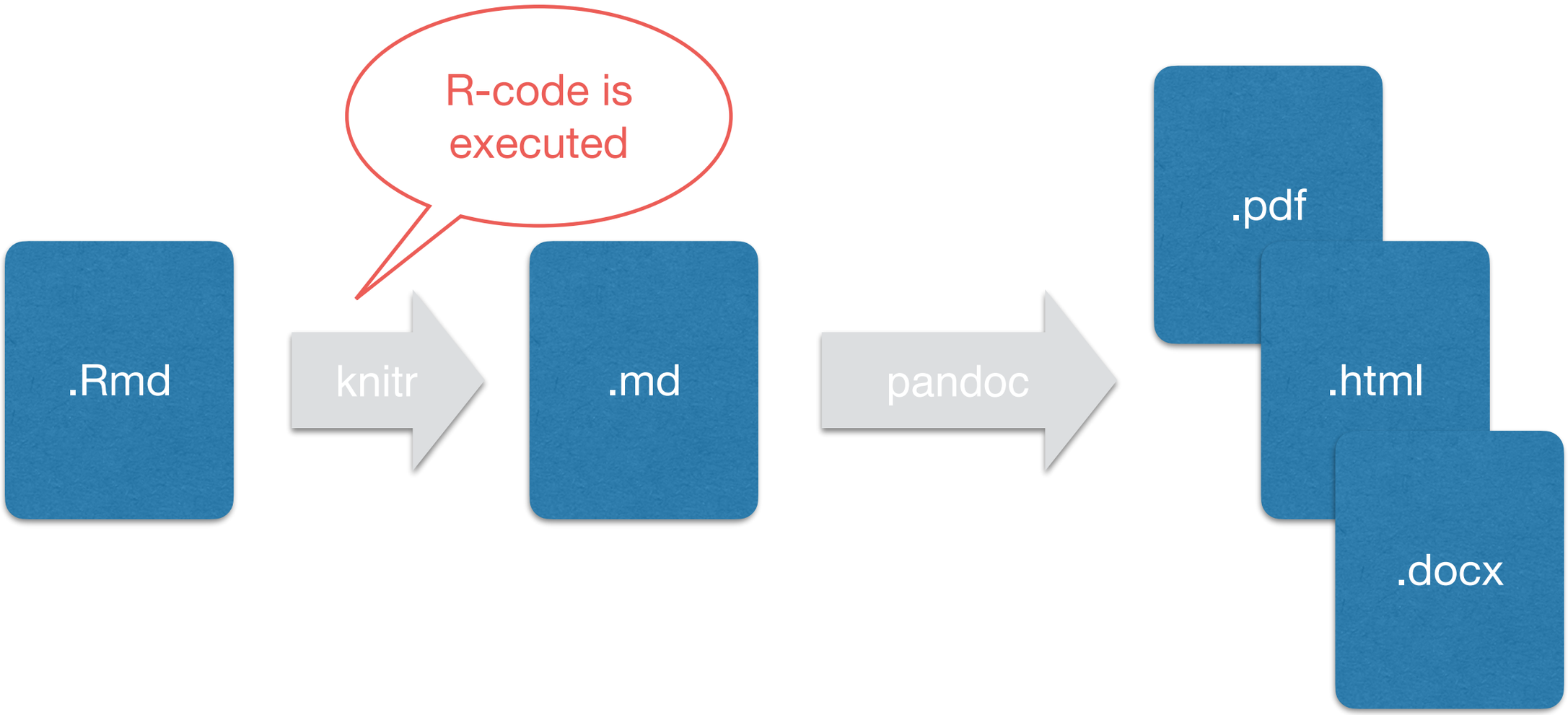
```
this is bash code
```

```
this is python code
```

Rmarkdown

Medizinische Fakultät Heidelberg

- With Rmarkdown, R-code parts can be included into the markdown document
- the R-code will be executed, the result integrated into markdown



Rmarkdown format

Medizinische Fakultät Heidelberg

````
---
title: "Project 01"
author: "Carl Herrmann"
date: "4/17/2019"
output:
  html_document:
    keep_md: yes
  pdf_document: default
---
# A Rmarkdown tutorial

This is a brief tutorial on how to use Rmarkdown to create dynamic documents

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_knit$set(root.dir='/Users/carlherrmann/Teaching/SS2019/DataAnalysis_4FS')
```

## Load the dataset

```{r read_data}
allDepMapData = readRDS('Data/depmap/DepMap19Q1_allData.RDS')
```

Now plot the distribution of the cell lines according to the tissue type

```{r plot_data}
freq = sort(table(allDepMapData$annotation$Primary.Disease))
par(las=2,mar=c(3,8,3,3));barplot(freq,horiz=TRUE, col='lightgrey')
```
````

header: set options

R code chunks

text in markdown

Rmarkdown chunk options

- Display options can be set for each chunk individually, or for all chunks at the beginning of the document

````
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(cache = TRUE)
```
````

valid for all chunks

- echo=TRUE : R-code is displayed in final document
- cache = TRUE : results of all chunks are cached

````
```{r plot_data,fig.height=12,fig.width=12}
freq = sort(table(allDepMapData$annotation$Primary.Disease))
par(las=2,mar=c(3,8,3,3));barplot(freq,horiz=FALSE, col='lightgrey')
```
````

valid for **this** chunks

- set height and width of output figure

Reference

- https://rmarkdown.rstudio.com/
- https://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf
-

Medizinische Fakultät Heidelberg

Conda

Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

(mini)conda

Medizinische Fakultät Heidelberg

- Conda is a tool to easily install software (for example python libraries)
- Conda can be used to create environments, which contain specific software
- These environment are independent of each other, i.e. do not interfere with each other
- Example
  - one environment with Python version 2.7 + libraries
  - one environment with Python 3 + libraries